DOCUMENT RESUME

ED 163 039 TM 008 047

AUTHOR Benson, Jeri; And Others
TITLE A Comparison of Three Types of Item Analysis in Test Development Using Classical and Latent Trait Methods.
PUB DATE [Mar 78]
NOTE 52p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978)

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Cognitive Tests; Comparative Analysis; *Efficiency; *Factor Analysis; *Item Analysis; Item Banks; Senior High Schools; *Test Construction; Test Items; *Test Reliability; Verbal Ability
IDENTIFIERS *Classical Test Theory; Latent Trait Model; *Rasch Model

ABSTRACT

The precision and efficiency of a cognitive test constructed by three different methods of item analysis was compared, using the verbal aptitude subtest of the Florida Twelfth Grade Test. Classical item analysis, factor analysis and the Rasch logistic model were used in the construction of 15 and 30 item subtests and replicated for samples of 250, 500, and 995 high school seniors. The study was designed in three phases: (1) item selection: (2) double cross-validation of the selected items: and (3) statistical analyses of the test and item characteristics. The results of the analysis showed that there were no apparent differences in the types of tests produced by the three methods of item analysis with regard to the precision of measurement. However, in terms of test efficiency, the results indicated substantive differences in the tests produced by the three methods of item analysis for students with varying ability levels. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED163039

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A COMPARISON OF THREE TYPES OF ITEM ANALYSIS IN TEST DEVELOPMENT USING CLASSICAL AND LATENT TRAIT METHODS

Jeri Benson
University of Southern California

Linda M. Crocker
University of Florida

William B. Ware
University of Florida

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Linda Crocker
William Ware

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada, 1978.

PRINTED IN U.S.A.

TM008 047

ABSTRACT

A COMPARISON OF THREE TYPES OF ITEM ANALYSIS IN TEST DEVELOPMENT USING CLASSICAL AND LATENT TRAIT METHODS

Jeri Benson
University of Southern California

Linda Crocker
University of Florida

William B. Ware
University of Florida

This study was designed to empirically compare the precision and efficiency of a cognitive test constructed by three different methods of item analysis. Classical item analysis, factor analysis and the Rasch logistic model were used in the construction of 15 and 30 item subtests; replicated for samples of 250, 500, and 995 examinees. The study was designed in three phases: (a) item selection, (b) double cross-validation of the selected items, and (c) statistical analyses of the test and item characteristics. The results of the analysis showed that there were no apparent differences in the types of tests produced by the three methods of item analysis with regard to the precision of measurement. However, in terms of test efficiency, the results indicated substantive differences in the tests produced by the three methods of item analysis for students with varying ability levels.

The systematic approach to test development was initiated by Binet & Simon in 1916. Since that time psychometricians have been concerned with the extent to which accurate measurement of a person's "ability" is possible. The accruracy of the measurement, or test precision is determined by the quality of the items in the test. Through the application of item analysis procedures, test constructors are able to obtain quantitative, objective information useful in developing and judging the quality of a test.

Classical test theory forms the basis for one method of test development. An integral part of the development of tests based on the classical model is the utilization of classical item analysis. Classical item analysis is a procedure to obtain a description of the statistical characteristics of each item of the test. This approach requires identification of single items which provide maximum discrimination between individuals on the latent trait being measured. Theoretically, selecting items which have high correlations with total test score will result in a discriminating test which is homogenous with respect to the latent trait. Therefore classical item analysis is an aid to developing internally consistent tests.

An alternative method of test development in classical test theory is factor analysis. Factor analysis is a more complex test development procedure than classical item analysis. It is a statistical technique that takes into account the item correlation with all other items in the test simultaneously. Groups of similar items tend to cluster together and comprise the latent traits (factors) underlying the test. Classical analysis then, can be viewed as a unidimensional basis for item analysis, less sophisticated than the multidimensional procedure of factor analysis. However, classical item analysis and factor analysis share a basic weakness, that of being sample dependent. Critics of classical test theory contend that the item statistics may vary when the examinee group changes; item statistics may also vary if a different set of items from the same domain is used with the same examinee group (Hambleton & Cook, 1977; Wright, 1968). Thus the selection of the final set of test items will be sample dependent.

Until recently, classical item analysis and factor analysis were the only techniques described in measurement texts for use in test development (Baker, 1977). However, since the publication of Lord and Novick's Statistical Theories of Mental Test Scores (1968) considerable attention has been

directed toward the field of latent trait theory as a new area in test development. Proponents of latent trait theory claim that the advantages over classical test theory are twofold: (a) theoretically it provides item parameters which are invariant across examinee samples which will differ with respect to the latent trait, and (b) it provides item characteristic curves that give insight into how specific items discriminate between students of varying abilities (Lord & Novick, 1968; Wright, 1968).

Several latent trait models have been developed; however this study was concerned only with the one-parameter logistic Rasch model (Rasch, 1960; 1966). The Rasch model was chosen because it is the most parsimonious of the latent trait mdels and has recently been used in the development and equating of tests (Rentz & Bashaw, 1977; Woodcock, 1977). In addition, studies have been conducted to verify the invariant properties of tests constructed using the Rasch model (Tinsley & Dawis, 1975; Whitely & Davis, 1974; Wright, 1968).

A review of the literature in the area of test development and item analysis, revealed that there have been no comparative studies involving all three item analysis techniques using the same data base. Also missing from the

literature were comparative studies of the three techniques across ability levels on measures of test efficency described by Lord (1974a; 1974b). Precision, as measured by internal consistency, is an overall estimate of a test's homogeniety, but provides no information on how the test as a whole discriminates for the various ability groups taking the test. For that reason measures of test efficiency were incorporated into the study. Test efficiency provides information on the effectiveness of one test over another as a function of ability level.

Wright (1977) has suggested that using the Rasch model with unweighted scores would be superior in estimating person ability than any other latent trait procedure. However, does this supposition extend to classical procedures, e.g., classical item analysis, and factor analysis? The purpose of this study was to compare the precision and efficiency of cognitive tests constucted by three methods of item analysis from a common item and examinee population.

METHOD

Subjects

A population of 78,751 high school seniors was available from 435 high schools throughout the state of Florida. From this population a 1 in 15 systematic sample of 5,250 students was chosen (Mendenhall, Ott, & Schaeffer, 1971). A systematic sample was chosen to ensure samples from every high school in the state.

The data file was edited to remove those students who either answered all items correctly or incorrectly. The rationale for this procedure was that the Rasch model cannot calibrate items when a person has a perfect score or the alternative, when a person has no items correct (Wright, 1977). Through the editing procedure 15 students were removed, thus the available sample size was 5,235. Because such a small number of students was removed, it seems unlikely that the elimination of these subjects would bias the results in favor of any of the three item analytic techniques.

Instruments

The instrument selected for use in this study was the Verbal Aptitude subtest of the Florida Twelfth Grade Test, developed by the Educational Testing Service. It is a statewide assessment battery which has been administered every year since 1935 (Benson, 1975). The Verbal Aptitude subtest is comprised of 50 verbal anagolies, in a multiple choice format, from which a single score based on the number of items correct is reported.

This particular instrument was selected for three reasons. First, it is a cognitive measure of verbal ability and much of classical test theory has been built upon tests in the cognitive domain. Second, it is similar to and hence representative of other national aptitude tests used for college admissions. Third, the items on college admissions tests have been written by experts, and each subtest is intended to be unidimensional. Thus, the assumptions from both classical and latent trait theory should be met.

Design

An empirical study was designed to compare the effects of the three methods of items analysis on test development for different sample sizes. The sample of 5,235 was divided into nine systematic samples as shown in Table 1.

Put Table 1 here

The purpose of obtaining the three separate samples for the three item analysis groups was to ensure that each item analytic technique used an independent sample, so that tests of statistical significance could be performed.

For each sample size two tests were developed by each item analytic procedure, a 15 item test, and a 30 item test. The following variables were obtained for each test:

- internal consistency
- standard error of measurement
- item difficulty
- item discrimination
- test efficiency curves

The independent variables were sample size and item analytic procedure.

The item data were analyzed in three phases: (a) selection of the items, (b) computation of item and test statistics for selected items on double cross-validation samples, and (c) statistical analyses of the item charac-

teristics of each test.

Procedure

Item selection. The three independent samples, within each of the three groups of students (N = 250, N = 500, N = 995), were submitted to one of the three item analytic procedures (in accordance with Table 1) in order to select a specified number ot items, e.g., the "best" 15 and 30 items.

Under classified item analysis, the best 15 and 30 items were selected based on the highest item biserial correlation with total test score. For factor analysis, the best 15 and 30 items were selected based on the highest item loadings on the first unrotated principal component. The selection of the items based on the Rasch model was accomplished in two stages. First, in order to check the assumption of a unidimensional test, a principal components analysis was run. Items were selected with loadings between .39 and .79 on the first unrotated factor, to hold the discrimination index of the items constant. Hambleton and Traub (1971) have shown that the efficiency of a test developed using the Rasch model will remain very high (over 95 percent) when the range on the discrimination index was held between .39 and .79. Second, the items selected from the principal

components solution using the above criteria were submitted to a Rasch analysis using the BICAL program (Wright & Mead, 1976). The best 15 and 30 items were selected based upon the item's mean square fit to the Rasch model.

Double cross-validation. A double cross-validation program (Mosier, 1951) was used to obtain item parameter estimates of 15 and 30 items selected by each item analytic technique for each sample size. This procedure ensured that estimates of parameters would be based upon a different sample of subjects than the sample used to originally identify the best items. Item analytic technique was reassigned, using a latin square (Cochran & Cox, 1957, p. 121) to a different sample within three groups. The double cross validation design is shown in Table 2.

Put Table 2 Here

The best 15 and 30 items selected by each item analytic procedure in the first phase of the study were submitted to a common item analysis program, GITAP (Baker & Martin, 1970) from which four of the dependent variables in the study were obtained: (a) indices of internal consistency as measured the analysis of variance procedure (Hoyt, 1941), (b) the

standard error of measurement, (c) item difficulty, and (d) item biserial correlations.

Statistical Analysis. The third phase of the study focused on obtaining measures of statistical significance for three of the dependent variables: internal consistency, item difficulty, biserial correlations.

The internal consistency estimates from each test were compared to the projected population value for tests of similar length via confidence intervals as suggested by Feldt (1965). (Projected population values were obtained using the Spearman-Brown Prophecy Formula).

Item difficulty and biserial correlations were submitted to four separate two-way analyses of variance for the 15 and 30 item tests. The two factors were sample size and item analytic procedure. The biserial correlations were transformed to an interval scale of measurement using a linear function of z suggested by Davis (1946) prior to analysis. Using Davis' transformation the range of the biserial correlations was 1 to 100. For comparison, in separate analyses actual biserial correlations were also submitted to the analysis of variance procedure. If statistical significance was observed in any of the analyses, with alpha equal to .05, Tukey's HSD post hoc procedure (Kirk,

(1968) was employed to determine which item analytic technique(s) resulted in the significant finding.

The criteria for evaluating the standard error of measurement produced by each item analytic technique will be set at a "meaningful" difference of > 1.00. This value was arbitrarily chosen prior to running any analyses.

In additon to the above tests, a measure of test efficiency was compared for the three 30 item tests, on the sample of 995 students.

Lord (1974a) has described a procedure to compare the relative efficiency of one test over another at different ability levels. If two tests to be compared vary in difficulty, then the relative efficiency of each will usually be different at different ability levels (Lord, 1974b; 1977). In classical test theory it is common to compare two tests that measure the same ability in terms of their reliability coefficients, but this only gives a single overall comparison. The formula developed by Lord for relative efficiency provides a more precise way of comparing two tests that measure the same ability. The formula for approximating relative efficiency is (Lord, 1974b, p. 248):

$$\text{R.E. } (y, x) = \frac{n_y}{n_x} \cdot \frac{x(n_x - x) f_x^2}{y\,(n_y - y)\, f_y^2} \quad (1)$$

where R.E. denotes the relative efficiency of y compared to x, n_y and n_x denote the number of items in the two tests, x and y are the number-right scores having the same percentile rank, and f_x^2 and f_y^2 are the squared observed frequencies of x and y. Lord has suggested that formula 1 only be used with a large sample of examinees and tests that are not extremely short, hence this comparison was restricted to the case where N = 995 and the 30 item test.

Three relative efficiency comparisons using the 30 item tests were made: (a) the test based on factor analysis was compared to the test based on classical item analysis, (b) the test based on the Rasch analysis was compared to the test based on classical item analysis, and (c) the test based on the Rasch analysis was compared to the test based on factor analysis.

RESULTS

The study was designed to compare empirically the precision and efficiency of tests developed using three item analytic techniques: classical item analysis, factor analysis and the Rasch model. The data regarding the precision of the 15 and 30 item test will be presented followed by the data on the efficiency of the 30 item test.

Comparison of the 15 Item Tests on Precision

The descriptive statistics based on the double cross-validation samples for the 15 item tests has been reported in Table 3.

Put Table 3 Here

The observed internal consistency estimates were tested for significance using confidence intervals described by Feldt (1965), to see if they were statistically different from the internal consistency estimate that would be obtained based on random item deletion.

The internal consistency estimate for the population based on the original 50 item subtest was .88 (Florida Twelfth Grade Testing Program, 1975). By applying the Spearman-Brown Prophecy Formula the projected population

internal consistency estimate for a 15 item test was found to be .687. The value .687 was the expected internal consistency estimate if 35 of the 50 items were randomly deleted.

The confidence intervals for the observed internal consistency estimates for each procedure have been presented in Table 4.

Put Table 4 Here

Of the nine confidence intervals, eight were found to be statistically different from the projected population estimate at the 95 percent level of confidence (the exception, N = 500 for the Rasch developed test). The three procedures tended to produce tests that were more precise in terms of internal consistency than would have been found if the items were randomly deleted.

The data reported in Table 3 indicated that the standard error of measurement for the 15 item tests based on the Rasch model were consistently larger than the standard error of measurement of the tests developed from classical item analysis and factor analysis for each sample test size. However, these differences were not meaningful

in that the difference did not equal or exceed 1.00 for any of the three procedures. Discriminations were tested for statistical significance in separate two-way analyses of variance to determine if there were differences in the types of items retained by each item analytic method. For item difficulty, no significant differences were found for item anlytic technique, sample size, or their interaction, F (2, 126) = 2.57, $p > .05$; F (2, 126) = .45, $p > .05$, F (4, 126) = .33, $p > .05$ respectively.

For the analysis of variance performed on the transformed biserial correlations of a significant F ratio was observed for the factor of item analytic technique, F (2, 126) = 14.862, $p < .05$. No significant differences were observed for sample size or the interaction of sample size and item analytic technique for the transformed biserial correlations, F (2, 126) = .30, $p > .05$; F (4, 126) = 1.16, $p > .05$ respectively. (When the actual biserial correlations were tested in the two-way analysis of variance design similar F ratios were observed.) The means standard deviations, and ranges of the item difficulties and transformed biserial correlations based upon the 15 item tests have been presented in Tables 5 and 6.

Put Tables 5 & 6 Here

Post hoc comparisons were made to determine which of the three item analytic procedures based upon their means contributed to the significant F ratio for the transformed biserial correlations for the 15 item tests. The mean transformed biserial correlation from the Rasch developed test was significantly lower than the mean biserial correlations from the tests developed by classical item analysis and factor analysis (Tukey's HSD value was 6.88, $p < .01$).

Comparison of the 30 Item Tests on Precision

The descriptive statistics based on the double cross-validation samples for the 30 item tests have been presented in Table 7.

Put Table 7 Here

The observed internal consistency estimates were tested for significance, using confidence intervals described in the previous section, to see if they were statistically different from the interval consistency estimate for the

population using random item deletion.

The projected population interval consistency estimate for a 30 item test was found to be .814 (via the Spearman-Brown Prophecy Formula). This value indicated the expected interval consistency if 20 of the 50 items were randomly deleted.

Based on the observed interval consistency estimates reported in Table 7, confidence intervals were generated for each item analytic procedure and have been presented in Table 8.

Put Table 8 Here

Of the nine confidence intervals, seven were found to be statistically different from the projected population estimate at a confidence level of 95 percent (the exception, N = 250, for the classical and Rasch developed test). Generally, the three procedures tended to produce tests that were more precise in terms of interval consistency than would have been found if the items were randomly deleted.

The standard error of measurement reported in Table 7 was essentially the same for the three methods of item analysis across the varying sample sizes.

Two-way analyses of variance were run on mean item difficulties and discriminations for the 30 item tests. No significant differences were observed for item difficulty for the independent variables of item analytic technique, sample size, or their interaction, F (2, 261) = .46, $p > .05$; F (2, 261) = .27, $p > .05$; F (4, 261) = .24 $p > .05$ respectively.

No significant differences were observed for the transformed biserial correlations for the independent variables of item analytic technique, sample size, or their interaction, F (2, 261) = .97, $p > .05$; F (2, 261) = .74, $p > .05$; F (4, 261) = .48, $p > .05$ respectively. (When the actual biserial correlations were tested similar F ratios were observed.) The means, standard deviations, and ranges of the item difficulties and transformed biserial correlations based upon the 30 item tests have been presented in Tables 9 and 10 respectively.

Put Tables 9 & 10 Here

Comparison of the 30 Item Tests on Efficiency

Three relative efficiency comparisons were made using the 30 item tests based on the sample of 995 examinees. The

three comparison were: (a) the test developed from factor analysis was compared to the test developed by classical item analysis, (b) the test developed by Rasch analysis was compared to the test developed by classical item analysis, and (c) the test developed Rasch analysis was compared to the factor analytically developed test.

The efficiency curves for the three comparisons were shown in Figure 1.

Put Figure 1 Here

The relative efficiency value was plotted on the ordinate, while the percentile rank (student ability level)was plotted along the abscissa. A relative efficiency of 1.00 would indicate that the tests were equally efficient.

A careful examination of formula 1 for relative efficiency indicated that when $n_x=n_y$ and $x=y$, that it was the number of examinees at the specified ability level(f_x^2 and f_y^2) that determined the efficiency of the test. That is, the fewer examinees observed at a particular percentile rank the better the test discriminates at that percentile rank. Therefore, test efficiency was equated with the level of discrimination the test was able to make between examinees, at various scores or percentile ranks.

The test developed by factor analysis was more efficient for the lower tenth of the pupils when compared to the test developed from classical item analysis. Both the tests were about equally efficient for the middle ability groups and high ability groups.

The Rasch developed test was more efficient than the test based on classical item analysis for average to high ability students (4oth-90th percentile rank). However, it was less efficient than the classical item analysis test for students with very low or very high abilities (1st-20th percentile rank and 98th percentile rank).

When compared to the factorially developed test, the Rasch test was again efficient for students of average to high abilities (50th-90th percentile rank). The factorially developed test appeared more efficient for the very low and very high ability students (1st-20th percentile rank and 98th percentile rank).

DISCUSSION AND CONCLUSIONS

This study was conducted to determine which of the three item analytic procedures (classical item analysis, factor analysis, and the Rasch model) might produce the superior test in terms of the precision and the efficiency

of measurement from a common item and examinee population. The discussion of the results then has been focused in two major areas: (a) the precision of the tests, and (b) the efficiency of the tests produced by the three methods of item analysis.

The Precision of the Tests Produced by the Three Methods of Item Analysis

Each of the three items analytic techniques was applied to an independent sample to select the best 15 and 30 items.

To test whether any of the three methods produced tests with greater internal consistency than a test created by random item deletion, the internal consistency estimates were compared to the projected internal consistency value for the population by using confidence intervals as suggested by Feldt (1965). In order for a given sample internal consistency estimate to be significant, the population value could not be included in the confidence interval generated around that sample value.

For the 15 and 30 item tests, 18 confidence intervals were generated for the observed internal consistency estimates, and 15 were found to be statistically significant at a 95 percent level of confidence. Thus, 15 observed internal consistency estimates were superior to the projected population value, where items were merely deleted at random.

Meaningful difference in the standard error of measurement were defined to be > 1.00, but none of the methods produced tests with standard errors of measurement that differed by that much for either the 15 or 30 item tests.

The two-way analysis of variance revealed no significant differences in the types of items selected for the 15 and 30 item tests with regard to item difficulty. Therefore, each item analytic technique tended to select items which had similar item difficulties on the average.

In this study, the item discriminations were measured by biserial correlations. The biserial correlations were transformed to an interval scale of measurement, and the 15 and 30 item tests were submitted to separate two-way analyses of variance. For the 15 item tests, a significant F ratio ($p < .05$) was observed for the main effect of item analytic technique. The mean transformed biserial correlation for the 15 item tests were: 44 for the Rasch test, 54 for the classical item analysis test, and 54 for the factorially developed test. (The actual mean biserial correlations for the three tests corresponding to the transformed biserial correlation were: .62, .71, .71 respectively). Tukey's HSD test revealed that the items selected by the Rasch procedure had significantly lower average biserial correlations than the items

selected by factor analysis and classical item analysis. It should be noted that this finding was probably due to the fact that the range of the biserial correlations was restricted to .39 - .79 on the items selected for the Rasch calibration. This was necessary to meet the assumption of equal item discriminations.

However, when the test length was increased to 30 items, no significant F ratios were observed for the variable of item discrimination. The difference in these two findings for the 15 and 30 item tests can be explained by the way the 15 and 30 item tests were constructed. The 15 item test was made up of 15 items with the highest biserial correlations. The 30 item test was made up of the above 15 items and an additional set of 15 items with the next highest biserial correlations. The addition of 15 more items meant that their average biserial correlation was something less than the original 15 items. Hence, the mean biserial correlations were reduced for the longer 30 item tests (Table 10).

From the data presented for each of the four areas above, it was concluded that each of the three item analytic techniques tended to produce tests that were really no different in terms of the precision of measurement.

Thus, the question to consider now is: Should

practitioners on the field of measurement spend their time learning to use the Rasch model to develop cognitive norm-referenced tests knowing the extra work and sophistication of knowledge required to effectively use the Rasch procedures? With the criterion of internal consistency as a measure of test superiority, it appeared from this study that time spent factorially developing test, or if computer facilities were not available, the use of classical item analysis procedures seem more than adequate for good test construction.

However, it must be remembered that internal consistency may not be a fair and sufficient criterion. Internal consistency is an integral part of classical test theory and may be biased since it was derived from the classical model. Following that reasoning, Whitely and Dawis (1974) have commented on the precision of tests developed using classical item analysis and Rasch analysis. They stated that if the goal of item selection was to develop fixed-content tests, then the classical techniques of item selection will yield the more precise test since precision is specific to the trait distribution in a given test. Whitely and Dawis indicated that the strength of the Rasch analysis was in the individualized selection of items, as in tailored testing, rather than construction of fixed-

content tests.

In considering the above situation, Lord (1974b) has stated that internal consistency is an overall estimate of a test's homogeneity, but provides no information on how the test as a whole discriminates for the various ability groups taking the test. Thus, the three techniques of test development were compared using an additional criterion, relative efficiency.

The Efficiency of the Test Produced by the Three Methods of Item Analysis

The relative efficiency formula (Lord, 1974a, 1974b) was not derived for any specific test development theory, therefore, relative efficiency estimates should be applicable to any test development technique. Lord (1974b) suggested his formula may not work well for extremely short tests and that it should only be used on large samples of examinees. Thus, in the present study, only the 30 item tests were compared using the sample of 995 examinees.

Generally, the results indicated that the Rasch test was superior to the two tests based on classical test theory for students of average to high ability The two test based on classical test theory, however, were superior in efficency to the Rasch developed test for very low and very high ability students (Figure 1). Test efficiency has been

defined as a measure of how well a test is able to discriminate between examinees of varying abilities. Therefore, the test constructor must ask himself, for which segment(s) of the examinee population is the test intended to discriminate?

In this study, the test under consideration was a verbal aptitude college admissions test. Usually college admissions officers are interested in selecting students who will be sucessful once admitted to college. The examinees who score very high on college admissions tests will generally be admitted to college without any question. Thus, it is less important to be able to discriminate among very high scoring examinees than to discriminate among students who score near the mean or in the upper middle range on a college admissions test. For these students it is difficult to decide who should be admitted and who should be denied admittance. If it is known that the admissions test discriminates very well for average to high ability students, then the reliability of the selection process based on test scores should be increased. The data in this study, therefore, indicate that the test based on the Rasch analysis would be most efficient for selecting the average to high ability students for admission to college.

Lord (1968) has illustrated a very important feature of test information and relative efficiency curves. He has shown that the contribution of each item to a test is independent of all other items. Thus, when information curves are available on a pool of items they can be added to a test to achieve a prespecified information or relative efficiency curve for any subpopulation of examinees (Lord, 1968). Therefore, by using measures such as relative efficiency and information curves psychometricians are able to develop very discriminating tests for any segment of the population.

It has been suggested that because efficiency and test information curves are a function of ability, these estimates ought to replace the use of classical reliability estimates and the standard error of measurement in test score information (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1977). This suggestion certainly deserves some consideration in light of the present study where it was shown that the three methods of item analysis produced similar tests in terms of precision, but the three methods produced very different tests in terms of efficiency. Today, with the increasing use of computers in test construction, perhaps the more meaningful question to be asked by psychometricians is: For which ability group is the test superior? Only test

information curves and measures of relative efficiency can answer that question.

Implications for Future Research

The results of this empirical study revealed that tests developed using classical test theory, in spite of its inherent weaknesses, were no different with respect to the precision of measurement than tests developed using one of the latent trait models, the one-parameter Rasch model. Comparisons of relative efficiency for the 30 item tests showed that the tests based on classical test theory were superior to the Rasch developed test for very low and very high scoring examinees, and the Rasch developed test was more efficient for average to high scoring examinees on the verbal aptitude college admissions subtest used in this study.

However, only one of the four latent trait models, the Rasch model, was used in this comparative study of test development techniques. Perhaps it was the very nature of this simple model that resulted in the development of equivalent tests when compared to the tests developed by classical item analysis and factor analysis in terms of precision. It may be that more technical two-and three-parameter logistic models would have produced tests comparable or

superior to those developed by classical test theory in terms of precision of measurement and overall relative efficiency.

If other latent trait models are to be considered in addition to the Rasch model, several points need to be evaluated. The Rasch model is the only latent trait model that provides for the direct calibration of items and abilities based on unweighted "number right" scoring (Wright, 1977). The two-and three-parameter models require a more complex scoring system where the item response is weighted in order to estimate the discrimination and guessing parameters. The weighting is an iterative process that may never converge or stablize unless arbitrary boundaries are established (Wright, 1977, p. 104). Because of this complex scoring system, the two-and three-parameter logistic models ar less efficient for parameter estimation than the one-par meter model in terms of computer time.

Additional areas for research may lead to actual comparisons of the content of the items selected by each of the three item analytic techniques. Davis (1951) and Cox (1965) have criticized the selection of items solely on statistical criteria. The use of statistical criteria alone, may result in changing the nature of the trait being

measured by deleting the items essential to adequate content coverage. Whitely and Dawis (1976) found this to be true in a study of verbal analogy items where the type of relationship and content of specific analogy items proved to be quite significant when studied in isolation.

REFERENCES

Baker, F.B. "Advances In Item Analysis." Review of Educational Research, 1977, 47, 151-178.

Baker, F.B., and Martin, T. Fortrap: A Fortran test analysis package. Occasional paper No. 10, Office of Research Consultation, College of Education, Michigan State University, 1970.

Benson, I.G. The Florida Twelfth Grade Testing Program: A Factor analytic study of the aptitude and achievement subtests. Unpublished Master of Arts in Education, thesis, University of Florida, 1975.

Binet, A., and Simon, T. The Development of Intelligence in Children. (E.S. Kite, trans.). Baltimore, Md.: Williams & Wilkins, 1916.

Cochran, W.G., and Cox, G.M. Experimental Designs. (2nd ed.) New York: John Wiley and Sons, Inc., 1957.

Cox, R.C. "Item Selection Techniques and Evaluation of Instructional Objectives. Journal of Educational Measurement, 1965, 2, 181-185.

Davis, E. Item Analysis Data: Their Computation, Interpretation and Use in Test Construction. Cambridge, Ma.: Harvard University, 1946.

Davis, E. "Item Selection Techniques." In E.F. Lindquist, ed. Educational Measurement. Washington, D.C.: American Council on Education, 1951.

Feldt, L.S. "The Approximate Sampling Distribution of Kuder-Richardson Reliability Coefficient Twenty." Psychometrika, 1965, 30, 357-365.

Florida Twelfth Grade Testing Program. (Report No. 1-75). Gainesville, Florida: Office of Instructional Resources, 1975.

Hambleton, R.K., and Traub, R.E. "Information Curves and Efficiency of Three Logistic Test Models." British Journal of Mathematical and Statistical Psychology. 1971, 24, 273-281.

Hambleton, R.K., and Cook, L. "Latent Trait Models and their Use in Analysis of Educational Test Data." Journal of Educational Measurement. 1977, 14, 75-96.

Hambleton, R.; Swaminathan, H.; Cook, L.; Eignor, D. and Gifford, J. Developments in Latent Trait Theory: A Review of Models, Technical Issues, and Applications. A paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Hoyt, C. "Test Reliability Estimated by Analysis of Variance." Psychometrika. 1941, 6, 153-160.

Kirk, R. Experimental Design: Procedures for the Behavioral Sciences. Belmont, Ca: Wadsworth, 1968.

Lord, F.M. "An Analysis of the Verbal Scholastic Aptitude Tests Using Birnbaum's Three-Parameter Logistic Model." Educational and Psychological Measurement. 1968, 28, 989-1020.

Lord, F.M. "The Relative Efficiency of Two Tests as a Function of Ability Level." Psychometrika. 1974, 39, 351-358 (a).

Lord, F.M. "Quick Estimates of the Relative Efficiency of Two Tests as a Function of Ability Level." Journal of Educational Measurement. 1974, 11, 247-254 (b).

Lord, F.M., and Novick, M. Statistical Theories of Mental Test Scores. Reading, Ma.: Addison-Wesley, 1968.

Marscuilo, L.A. Statistical Methods for Behavioral Science Research. New York: McGraw-Hill, 1971.

Mendenhall, W.; Ott, L. and Schaffer, R. Elementary Survey Sampling. Belmont, Ca.: Wadsworth Publishing Co., 1971.

Mosier, C.I. "Problems and Designs of Cross Validation." Educational and Psychological Measurement. 1951, 11, 5-11.

Rasch, G. Probablistic Models for Some Intelligence and Attainment Tests. Copenhagen, Denmark: Denmark's Paedagogiski Institute, 1960.

Rasch, G. "An Item Analysis Which Takes Individual Differences Into Account." British Journal of Mathematical and Statistical Psychology. 1966, 19, 49-57.

Rentz, R.R., and Bashaw, W.L. "The National Reference Scale for Reading: An Application of the Rasch Model." Journal of Educational Measurement. 1977, 14, 161-179.

Tinsley, H., and Dawis, R. "An Invstigation of the Rasch Simple Logistic Model: Sample-Free Item and Test Calibration." Educational and Psychological Measurement. 1975, 35, 325-339.

Whitely, S., and Dawis, R. "The Nature of Objectivity with the Rasch Model." Journal of Educational Measurement. 1974, 11, 163-178.

Whitely, S., and Dawis, R. "The Influence of the Context on Item Difficulty." Educational and Pscyhological Measurement. 1976, 36, 329-337.

Woodcock, R.W. Woodcock Reading Mastery Test. Circle Pines, Minn.: American Guidance Service, 1974.

Wright, B. "Sample-Free Test Calibration and Person Measurement." Proceedings of the 1967 Invitation Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B. "Solving Measurement Problems with the Rasch Model." Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. and Mead, R.J. BICAL: Calibrating rating scales with the Rasch Model. Research Memorandum No. 23. Chicago: Statistical Laboratory, University of Chicago, 1976.

TABLE 1

Systematic Sampling Design of the Study[a]
N = 5,235

Group	Sampling Procedure	Sample Number	Number Selected	Item Analytic Procedure	Total Sample Remaining
$Group_1$	1 in 20	1	250	Classical	4,985
	1 in 19	2	250	Factor Analysis	4,735
	1 in 18	3	250	Rasch	4,485
$Group_2$	1 in 8	4	500	Classical	3,985
	1 in 7	5	500	Factor Analysis	3,485
	1 in 6	6	500	Rasch	2,985
$Group_3$[b]	1 in 3	7	995	Classical	1,995
	1 in 2	8	995	Factor Analysis	995
	remaining	9	995	Rasch	0

[a] The sampling procedure was randomly assigned to item analytic technique in group and the same pattern carried out for group and group .

[b] Those subjects editied from the data file were removed equally from group , hence the reduced sample file

TABLE 2

Double Cross-Validation Design of the Study

Group	Sample Number[a]	Number Selected	Item Analytic Procedure	Double Cross-Validation Procedure[b]
$Group_1$	1	250	Classical	Factor Analysis
	2	250	Factor Analysis	Rasch
	3	250	Rasch	Classical
$Group_2$	4	500	Classical	Rasch
	5	500	Factor Analysis	Classical
	6	500	Rasch	Factor Analysis
$Group_3$	7	995	Classical	Rasch
	8	995	Factor Analysis	Classical
	9	995	Rasch	Factor Analysis

[a] The sample number is the same as referred to in Table 1.

[b] Assignment to sample was based on a randomized 3 X 3 latin square procedure.

TABLE 3

DESCRIPTIVE STATISTICS FOR THE TEST COMPOSED OF THE 15 BEST ITEMS SELECTED BY EACH ITEM ANALYTIC PROCEDURE ACCORDING TO SAMPLE SIZE

	N = 250			N = 500			N = 995		
	Classical	Factor Analysis	Rasch	Classical	Factor Analysis	Rasch	Classical	Factor Analysis	Rasch
Test:									
Mean	9.80	10.40	9.41	9.27	10.21	8.88	10.10	9.72	8.76
Standard Deviation	3.41	3.41	3.24	3.69	3.50	3.18	3.63	3.62	3.22
Internal Consistency[a]	.790	.800	.754	.816	.811	.724	.826	.814	.735
Standard Error of Measurement	1.51	1.47	1.55	1.53	1.47	1.61	1.46	1.51	1.60
Difficulty:									
Mean	.65	.69	.63	.62	.68	.60	.67	.65	.58
Standard Deviation	.16	.14	.17	.15	.15	.16	.15	.14	.17
Range	.31-.89	.40-.92	.32-.90	.32-.83	.35-.92	.36-.89	.33-.91	.37-.90	.27-.81
Biserial[b]:									
Mean	50.93	56.13	46.53	53.53	53.53	41.20	56.33	53.80	42.80
Standard Deviation	14.52	17.05	8.94	12.45	8.20	7.95	9.19	9.44	7.26
Range	39-82	34-91	37-64	29-78	39-66	28-56	41-73	38-70	30-56

[a] Internal consistencies reported to three digits to observe slight differences.

[b] Based on linear transformation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1946)

TABLE 4

[a] Confidence Intervals for the Observed Internal Consistency Estimates Based on the 15 Item Tests According to Sample Size

Procedure	95% Confidence Interval N = 250	N = 500	N = 995
Classical	.748 - .828*	.792 - .838*	.810 - .841*
Factor Analysis	.760 - .837*	.786 - .854*	.797 - .831*
Rasch	.704 - .799*	.688 - .757	.711 - .759*

[a] The F values used in calculating the confidence intervals were obtained from Marisculo (1971).

* Statistical significance is indicated when the population internal consistency estimate is not included in the confidence interval generated for each observed internal consistency estimate. The projected population value as .687.

TABLE 5

15 Item Tests:
Descriptive Statistics for Item Difficulty
by Procedure and Sample Size

	Procedure			Sample Size		
	Classical	Factor Analysis	Rasch	250	500	995
Mean	.65	.67	.60	.66	.63	.63
Standard Deviation	.15	.14	.16	.15	.15	.16
Range	.31 - .91	.35 - .92	.27 - .90	.31 - .92	.32 - .92	.27 - .91

TABLE 6

15 Item Tests:
Descriptive Statistics for Item Discriminations[a]
by Procedure and Sample Size

	Procedure			Sample Size		
	Classical	Factor Analysis	Rasch	250	500	995
Mean	53.60	54.49	43.51	51.20	49.47	50.98
Standard Deviation	12.18	11.98	8.21	14.17	11.20	10.36
Range	29-82	34-91	28-64	34-91	28-78	30-73

[a]Based on transformed biserial correlations. The transformation was a linear transformation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1946).

TABLE 7

DESCRIPTIVE STATISTICS FOR THE TEST COMPOSED OF THE 30 BEST ITEMS SELECTED BY EACH ITEM ANALYTIC PROCEDURE ACCORDING TO SAMPLE SIZE

	N = 250			N = 500			N = 995		
	Classical	Factor Analysis	Rasch	Classical	Factor Analysis	Rasch	Classical	Factor Analysis	Rasch
Test;									
Mean	17.12	17.98	16.99	17.69	17.73	18.24	17.31	18.51	17.54
Standard Deviation	5.73	6.05	5.68	6.30	6.06	5.80	6.22	6.22	6.11
Internal Consistency[a]	.834	.854	.828	.863	.853	.838	.862	.861	.851
Standard Error of Measurement	2.29	2.27	2.32	2.29	2.27	2.30	2.27	2.28	2.31
Difficulty:									
Mean	.57	.60	.57	.59	.59	.61	.58	.62	.59
Standard Deviation	.19	.17	.18	.16	.17	.16	.17	.15	.16
Range	.23-.89	.29-.92	.23-.90	.32-.89	.31-.92	.24-.89	.21-.91	.35-.90	.27-.85
Biserial[b]:									
Mean	38.67	42.17	37.43	42.30	40.63	38.17	42.10	42.03	40.07
Standard Deviation	13.30	13.13	10.67	11.19	9.33	8.82	10.85	10.03	10.34
Range	9-70	10-68	13-57	24-68	21-58	23-55	25-66	28-61	24-64

[a] Internal consistencies reported to three digits to observe slight differences

[b] Based on linear trasnformation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1946).

TABLE 8

Confidence Intervals[a] for the Oberserved Internal Consistency Estimates Based on th 30 Item Tests According to Sample Size

	95% Confidence Interval		
	N = 250	N = 500	N = 955
Classical	.800-.865	.845-.880*	.850-.874*
Factor Analysis	.825-.881*	.834-.871*	.848-.873*
Rasch	.794-.860	.817-.858*	.838-.863*

[a]The F values used in calculationg the confidence intervals were obtained from Marisculo (1971).

*Statistical significance is observed when the population internal consistency estimate is not included in the confidence interval generated for each observed internal consistency estimate. The projected population value was .814.

TABLE 9

30 Item Tests: Descriptive Statistics for Item Difficulty by Procedure and Sample Size

	Procedure			Sample Size		
	Classical	Factor Analysis	Rasch	250	500	995
Mean	.58	.60	.59	.58	.60	.59
Standard Deviation	.17	.16	.17	.18	.16	.16
Range	.21-.91	.29-.92	.23-.90	.23-.92	.24-.92	.21-.91

TABLE 10

30 Item Tests: Descriptive Statistics for Item Discriminations[a] by Procedure and Sample Size

	Procedure			Sample Size		
	Classical	Factor Analysis	Rasch	250	500	995
Mean	41.02	41.61	38.56	39.42	40.37	41.40
Standard Deviation	11.82	10.85	9.93	12.45	9.87	10.34
Range	9-70	19-68	13-64	9-70	21-68	24-66

[a]Based on transformed biserial correlations. The transformation was a linear transfromation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1946).

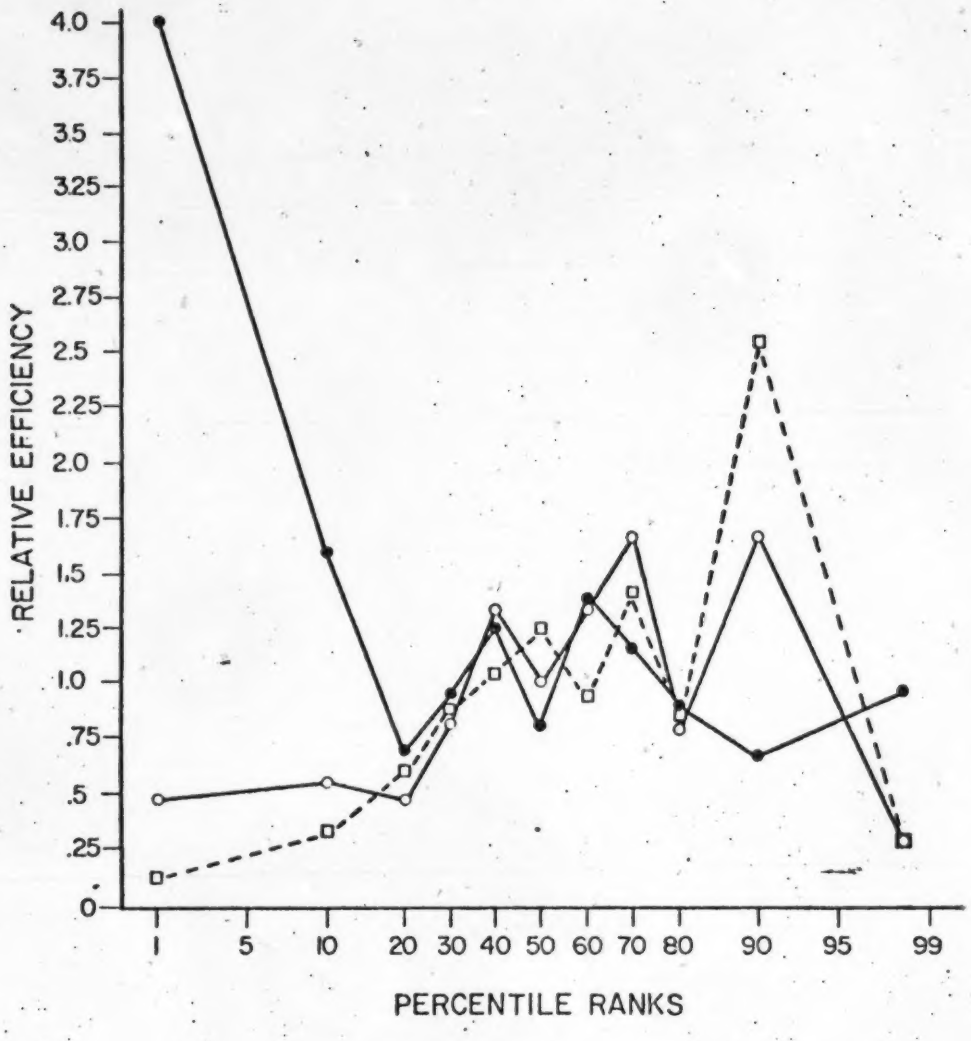

FIGURE 1. RELATIVE EFFICIENCY COMPARISONS FOR THE THREE 30 ITEM TESTS N=995.